ISSN 1000-0054 清华大学学报 (自然科学版) 2008 年 第 48 卷 第 S1 期 19/27
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2008, Vol. 48, No. S1 715-719

# 自适应权重的双模态情感识别

黄力行, 辛 乐, 赵礼悦, 陶建华

(中国科学院 自动化所, 模式识别国家重点实验室, 北京 100080)

**摘 要**: 情感识别是人机交互领域的重要问题之一。语音和脸部肌肉动作信息是用于情感识别的 2 个最重要的模态。该文认为, 在双模态情感识别中, 给不同的特征赋予不同的权值有利于充分利用双模态信息, 提出了一种基于 Boosting 算法的双模态信息融合方法, 它能够自适应地调整语音和人脸动作特征参数的权重, 从而达到更好的识别效果。实验表明, 该方法能够更好地区分易混淆的情感状态, 情感识别率达 84% 以上。

**关键词**: 双模态情感识别; Boosting 算法; 自适应权重

**中图分类号**: TP 3 **文献标识码**: A

**文章编号**: 1000-0054(2008)S1-0715-05

## Bimodal emotion recognition based on adaptive weights

HUANG Lixing, XIN Le, ZHAO Liyue, TAO Jianhua

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

**Abstract**: Emotion recognition is one of the most important issues in human-computer interactions (HCI). This paper describes a bimodal emotion recognition approach using a boosting-based framework to automatically determine the adaptive weights for audio and visual features. The system dynamically balances the importance of the audio and visual features at the feature level to obtain better performance. The tracking accuracy of the facial feature points is based on the traditional KLT algorithm integrated with the point distribution model (PDM) to guide analysis of the deformation of facial features. Experiments show the validity and effectiveness of the method, with a recognition rate over 84%.

**Key words**: bimodal emotion recognition; boosting; adaptive weights

近年来, 情感识别的研究工作[1-9]在人机交互领域中已经成为一个热点。过去很多的工作都是集中在如何通过单模态的信息[5,10-13], 如语音或者人脸表情, 得到当前对象的情感状态。仅仅通过单模态信息来识别情感有很多的局限性, 因为人类是通过多模态的方式表达情感信息的。最近, 基于多模态, 尤其是基于语音和人脸表情双模态的情感识别技术得到了很大的发展。

目前, 融合多模态信息的方法主要有2 种: 决策层的融合和特征层的融合。决策层的融合技术是先把各个模态的信息提取出来, 输入相应的分类器得到单模态识别结果, 然后用规则的方法将单模态的结果综合起来, 得到最终的识别结果; 特征层的融合方法则是将各个模态的信息提取出来, 将这些信息组成一个统一的特征向量, 然后再输入到分类器中, 得到最终的识别结果。这 2 种方法各有优缺点。决策层的融合技术考虑了不同模态对于情感识别重要性的不同, 如文[6]认为, 在识别不同情感的时候, 语音和人脸表情的重要性不同, 因此他们通过主观感知实验给语音和人脸表情信息赋予不同的权重。但是这种通过主观感知实验得到的权重能否应用到其他的情况下是值得怀疑的。特征层的融合技术更接近人类识别情感的过程, 能更好地利用统计

收稿日期: 2007-09-10
基金项目: 国家自然科学基金资助项目 (60575032);
国家“八六三”高技术项目 (2006AA01Z138)
作者简介: 黄力行(1984—), 男(汉), 江西, 硕士研究生。
通讯联系人: 陶建华, 副研究员, E-mail: jhtao@nlpr.ia.ac.cn

机器学习的技术。文[7]将语音和人脸表情的信息综合成一个特征向量,并使用支持向量机进行分类,得到最终的识别结果。但是这种方法没有考虑到识别不同情感时,不同模态重要性的不同,因此这种方法不能最大程度地发挥双模态融合的优势。

为了能将决策层和特征层融合的优点结合起来,本文提出了一种基于boosting的双模态融合方法。语音和人脸表情信息首先被融合到一个统一的特征向量中,然后再使用以分类回归树(classification and regression trees, CART)为弱分类器的强分类器,得到最终识别结果。在训练弱分类器的过程中,通过给每一个训练样本赋予不同的权重,自动调整不同特征在双模态融合过程中的重要性。实验表明,和以前的方法[6-7]相比,这种方法能够较好地区分易混淆的情感状态,得到更高的识别率。

## 1 双模态情感识别框架

### 1.1 系统框架

系统由3部分构成,如图1所示,分别是声学参数提取模块,人脸特征点参数提取模块和双模态特征向量分类模块。该分类模块将双模态特征向量分为中性、高兴、悲伤、愤怒、害怕和惊讶6种情感,它由一系列的分类回归树模型组成,能够在训练的过程中调整各个参数的重要性,从而获得更好的识别结果。

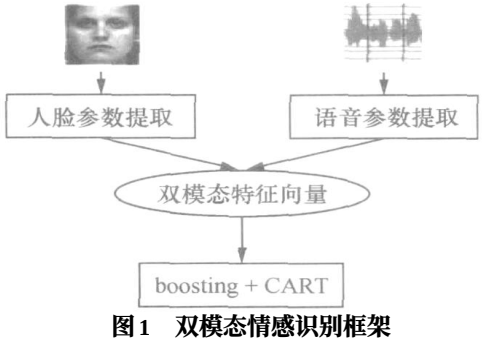


图1 双模态情感识别框架

### 1.2 语音参数提取

自动语音切分、基频提取、短时能量计算等语音信号处理算法已经成熟。通过对语音信号的预处理和特征提取,能够得到各种声学参数。前人的研究表明:在众多的语音参数中,时长、基频的范围、基频的最大值最小值、基频的均值、能量的均值等都是用于情感识别的较为有效的特征。为了强调重音的作用,文[14]又引入了基频最大值和最小值的位置、时长最大值和最小值的位置,详细分析了不同语音参数在情感识别中的重要性。结果显示,基频的均值、基频的最大值、基频的范围、能量的均值、时长的均值和基频最小值的位置是最为重要的语音参数。因此,本文在声学参数提取部分也使用了这些参数。

### 1.3 人脸参数提取

人脸参数提取基于人脸特征点的跟踪。考虑到跟踪算法的鲁棒性,这里选取的特征点都是在像素值上具有较明显梯度的点,如图2所示。

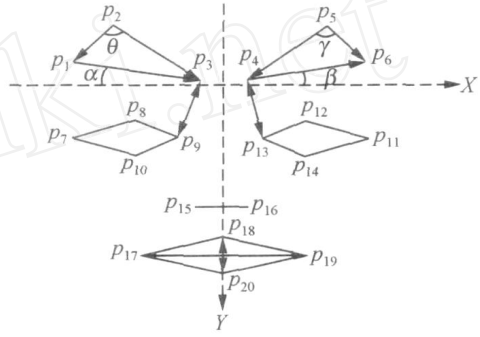


图2 人脸特征点及其几何参数

根据特征点跟踪结果,本文采用了如下特征作为人脸表情参数:

$$\begin{cases} \phi_1 = \frac{1}{2}(\theta + \gamma), \\ \phi_2 = \frac{1}{2}(\alpha + \beta), \\ d_1 = \frac{1}{2}(|\overrightarrow{p_3p_9}| + |\overrightarrow{p_4p_{13}}|), \\ d_2 = |\overrightarrow{p_{17}p_{19}}| \end{cases} \tag{1}$$

所选特征基本都位于上半脸,这是因为嘴部附近的运动很大程度上受到说话内容的影响。

## 2 基于Boosting的识别算法

通过语音参数提取和人脸参数提取模块,得到了双模态情感识别的训练数据。假设数据集$S = \{(x_i, y_i)\}_{i=1}^{\Pi}$,其中:$\Pi$是训练数据集的大小,$x_i$是双模态特征向量,$y_i$是情感类别。这里考虑的情感类别数为6,分别是中性、高兴、悲伤、害怕、惊讶和生气,即$y_i \in Y = \{0, 1, 2, 3, 4, 5\}$。

基于boosting的算法是在训练数据集上利用迭代的方法不断地产生弱分类器,然后将这些弱分类器线形的组合在一起,形成强分类器。本文中使用的弱分类器是CART模型。图3是强分类器的构

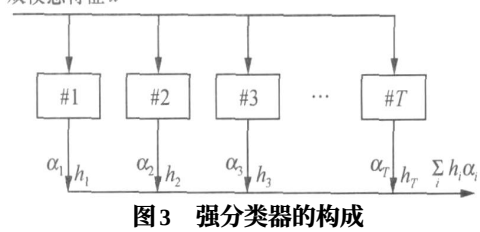


图3 强分类器的构成

成, $T$ 是迭代的次数。

对于每一个样本 $x_i$, 它的预测的类别是 $\hat{k} = \arg\max_k \{h_t(x_i, k)\}$, 其中 $k \in Y$, $h_t(x_i, k)$ 表示第 $t$ 个 CART 模型将样本 $x_i$ 预测为类别 $k$ 的概率。对于训练集中的每一个样本 $(x_i, y_i)$, 可以得到 2 组概率 $h_t(x_i, l_0)$ 和 $h_t(x_i, l_1)$, 其中 $l_0 \neq y_i$, $l_1 = y_i$。

Boosting 算法:

1) 给定 $M$ 个训练数据, 分别是 $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_M, y_M)$, 其中 $y_i \in Y$。

2) 初始化每个样本的权值

$$D_1(i, l_0, l_1) = \begin{cases} 1/(M \cdot |y_i| \cdot |Y - y_i|), & l_0 \neq y_i \text{ 和 } l_1 = y_i, \\ 0, & \text{其他} \end{cases} \tag{2}$$

其中: $|y_i|$ 是样本属于的类别的数目, $|Y - y_i|$ 是剩下的类别的数目。对于本问题而言, $|y_i| = 1$, $|Y - y_i| = 5$, 即每个样本有 5 个非零的权值。

3) 迭代 $t = 1, 2, 3, \cdots, T$。

a) 利用当前样本权值的分布训练弱分类器 $h_t$;

b) 计算当前弱分类器的权重 $\alpha_t \in R$;

c) 更新样本的权重

$$D_{i+1}(i, l_0, l_1) = \frac{D_i(i, l_0, l_1)\exp\{\alpha_t[h_t(x_i, l_0) - h_t(x_i, l_1)]/2\}}{Z_t}, \tag{3}$$

$Z_t$ 是归一化参数。

4) 输出最后的强分类器

$$f(x, k) = \sum_{t=1}^{T} \alpha_t h_t(x, k).$$

从式(3)可以看出, 对于那些在当前轮被错分的样本, 即 $h_t(x_i, l_0) < h_t(x_i, l_1)$, 它们的权重会增加, 这就会使得下一轮训练的弱分类器更加关注当前被错分的样本。弱分类器通过重采样的方式关注被错分的样本。假设当前轮的权重分布是 $D_t(k, l_0, l_1)$, 那么样本 $i$ 的权重是 $\sum_{l_0, l_1} D_t(i, l_0, l_1)$, 该样本在下一轮中出现的次数为

$$|x_i|_{t+1} = \frac{\sum_{l_0, l_1} D_t(i, l_0, l_1)}{\min_{j=\{1, 2, \cdots, M\}} \left\{ \sum_{l_0, l_1} D_t(i, l_0, l_1) \right\}}. \tag{4}$$

也就是说, 那些权重增加的样本会复制自己, 使之在下一轮的训练集中所占的比例增加。在这个重采样的训练集合上, 得到新的弱分类器 $h_{t+1}$。$\alpha_{t+1}$ 由式(5)、(6)可得:

$$r_{t+1} = \sum_{i, l_0, l_1} D_{t+1}(i, l_0, l_1)[h_{t+1}(x_i, l_1) - h_{t+1}(x_i, l_0)], \tag{5}$$

$$\alpha_{t+1} = \frac{1}{2}\ln\left(\frac{1 + r_{t+1}}{1 - r_{t+1}}\right). \tag{6}$$

当一个双模态特征向量输入最终的强分类器时, 预测的情感类别为 $f(x) = \arg\min_{k \in Y} \sum_{t=1} \alpha_t h_t(x, k)$。

## 3 实验和讨论

### 3.1 数据库

本文中用到的数据库有 2 个男声和 2 个女声, 分别录制了 6 种表情(中性、高兴、悲伤、害怕、惊讶和生气), 各 500 句双模态数据, 包括脸部表情和同步的语音数据。每一句有 6～12 个音节。因此, 双模态数据库中有 $500 \times 6 \times 4 = 12\,000$ 句话, 每个情感集中有 $4 \times 500 = 2\,000$ 个样本。对于每个说话人, 本文使用了 $300 \times 6$ 个样本进行训练, 在剩下的样本上测试。语音的采样率是 22 kHz, 图像的采样率是 30 帧/s。

### 3.2 不同特征的重要性

训练开始时, 每个样本的权重都相同。随着迭代的进行, boosting 算法能自动地调整样本的权重, 那些容易被错分的样本不断地增加自己在训练集中的比例, 如图 4 所示。

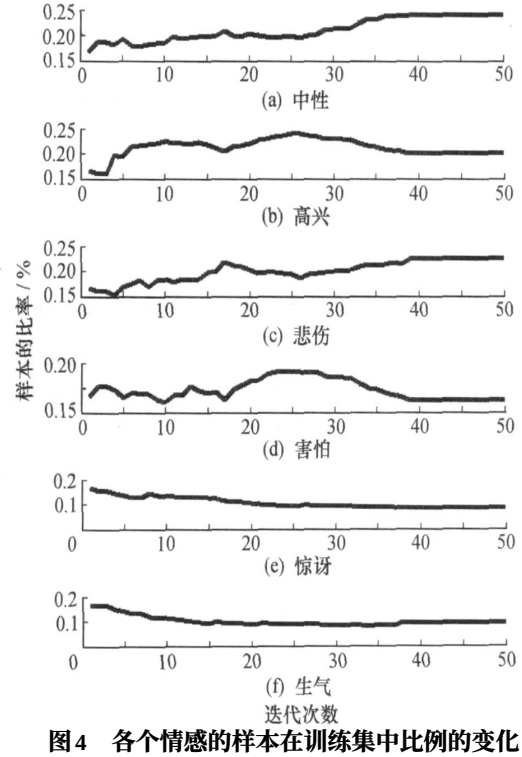


图 4 各个情感的样本在训练集中比例的变化

本文利用CART 模型作为弱分类器, 因此可以根据CART 模型的结构确定每个特征的重要性。由于CART 模型使用贪心算法, 每次从特征集合中选出最有效的特征作为分类的标准, 因此不难得出: 1) 一个特征在CART 模型中出现的次数越多, 这个特征的重要性就越高; 2) 特征的重要性随着决策树深度的增加而减少, 即树根处的特征最重要。特征的重要性

$$\mathrm{DOMINANCE}(f) = \sum_{f \in F} p^{h(f)}, \qquad (7)$$

其中: $F$ 包含了出现在树里的所有特征, $h(f)$ 是特征 $f$ 出现在树中的高度, $p$ 是介于0～1之间的常数。图5是各个特征的重要性随着迭代次数增加的变化。

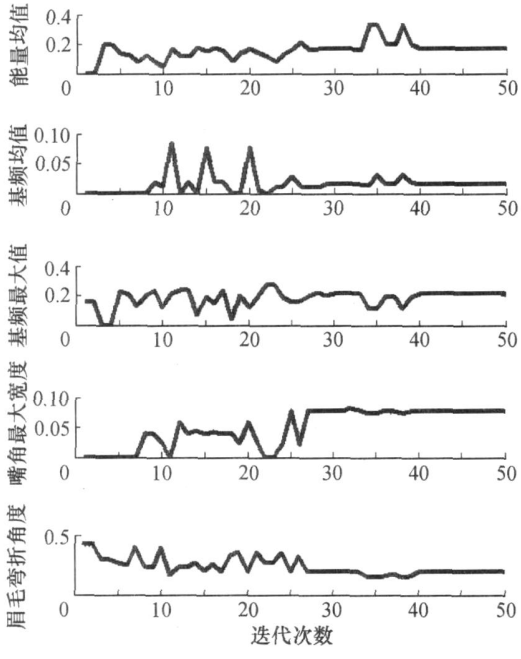


图5 特征重要性随着迭代进行的变化

通过对比图4和图5, 可以看出基于boosting 的迭代算法虽然直接改变的是训练样本的权重, 但是它内在地调整了不同特征的重要性, 也就是说这种算法能够根据训练样本自动地选择出最好的特征进行分类。例如, 随着高兴样本的增加(图4), 2个嘴角之间的宽度这个特征的重要性也随之增加(图5)。这也符合人类表现情感的方式。

### 3.3 对比实验数据

本文使用6×6混淆矩阵来评价情感识别算法的好坏。第 $i$ 行第 $j$ 列的元素表示真实情感状态是 $i$ 的样本被判别成是 $j$ 的比例。也就是说, 矩阵对角线上的值越大, 情感识别算法的效果就越好。表1和表2分别表示只利用语音信息或者人脸信息得到的判别结果。

表1 只利用语音特征的情感识别结果

%

| | 中性 | 高兴 | 悲伤 | 生气 | 害怕 | 惊讶 |
|---|---|---|---|---|---|---|
| 中性 | 70 | 12 | 4 | 0 | 8 | 6 |
| 高兴 | 10 | 78 | 2 | 2 | 0 | 8 |
| 悲伤 | 22 | 6 | 40 | 0 | 30 | 2 |
| 生气 | 0 | 0 | 0 | 94 | 0 | 6 |
| 害怕 | 20 | 8 | 32 | 0 | 40 | 0 |
| 惊讶 | 0 | 4 | 2 | 32 | 0 | 62 |

表2 只利用人脸特征的情感识别结果

%

| | 中性 | 高兴 | 悲伤 | 生气 | 害怕 | 惊讶 |
|---|---|---|---|---|---|---|
| 中性 | 72 | 0 | 26 | 0 | 2 | 0 |
| 高兴 | 40 | 40 | 0 | 20 | 0 | 0 |
| 悲伤 | 2 | 18 | 80 | 0 | 0 | 0 |
| 生气 | 0 | 0 | 10 | 78 | 0 | 12 |
| 害怕 | 18 | 0 | 0 | 0 | 50 | 32 |
| 惊讶 | 10 | 0 | 2 | 0 | 28 | 60 |

从上面的结果可以知道, 只用语音信息, 悲伤和害怕与惊讶和生气是2对容易混淆的情感; 只利用人脸信息, 中性和高兴与害怕和惊讶是2对容易混淆的情感。

表3是利用boosting 算法的双模态情感识别结果。可以清楚地看到, 由于双模态信息的互相补充, 使得原来容易混淆的情感对能够被分开, 从而大大提高了情感识别的准确性。

表3 利用双模态信息的情感识别结果

%

| | 中性 | 高兴 | 悲伤 | 生气 | 害怕 | 惊讶 |
|---|---|---|---|---|---|---|
| 中性 | 88 | 6 | 6 | 0 | 0 | 0 |
| 高兴 | 14 | 84 | 2 | 0 | 0 | 0 |
| 悲伤 | 0 | 2 | 85 | 0 | 13 | 0 |
| 生气 | 0 | 2 | 0 | 96 | 0 | 2 |
| 害怕 | 0 | 0 | 5 | 0 | 95 | 0 |
| 惊讶 | 0 | 1 | 0 | 6 | 0 | 93 |

为了证明自适应的调整特征的重要性有助于提高情感识别的准确率, 本文重复了文[6]和文[7]中的方法。这2种方法分别是基于决策层的融合和基于特征层的融合。表4和表5分别是这2种方法得到的混淆矩阵。

表4 基于决策层融合的情感识别结果

%

| | 中性 | 高兴 | 悲伤 | 生气 | 害怕 | 惊讶 |
|---|---|---|---|---|---|---|
| 中性 | 58 | 15 | 15 | 12 | 0 | 0 |
| 高兴 | 8 | 65 | 4 | 23 | 0 | 0 |
| 悲伤 | 8 | 0 | 75 | 3 | 14 | 0 |
| 生气 | 2 | 11 | 2 | 85 | 0 | 0 |
| 害怕 | 6 | 2 | 13 | 0 | 79 | 0 |
| 惊讶 | 0 | 0 | 0 | 10 | 4 | 86 |

表5 基于特征层融合(SVM)的情感识别结果

%

| | 中性 | 高兴 | 悲伤 | 生气 | 害怕 | 惊讶 |
|---|---|---|---|---|---|---|
| 中性 | 75 | 25 | 0 | 0 | 0 | 0 |
| 高兴 | 6 | 85 | 0 | 9 | 0 | 0 |
| 悲伤 | 3 | 0 | 80 | 0 | 17 | 0 |
| 生气 | 0 | 0 | 0 | 90 | 0 | 10 |
| 害怕 | 0 | 0 | 7 | 0 | 88 | 5 |
| 惊讶 | 0 | 0 | 0 | 5 | 0 | 95 |

对比表3和表4的结果, 可以清楚地看到, 基于boosting的方法好于单纯的决策层融合的方法。这是因为不同的人有不同的表达习惯, 不同的情感也有不同的表达方式, 用一个固定的规则去识别不同人的各种情感显然是不合适的。而基于boosting的算法能够自动地分析不同特征的重要性, 这种数据驱动的方法显然要优于基于规则的融合方法。

而表3和表5的对比发现, 基于boosting的方法略好于用SVM的方法。尽管SVM是一个能力很强的分类器, 但是由于这种特征融合的算法仅仅是简单地把特征向量放在了一起而没有考虑不同特征的重要性, 使得它的识别结果仍然低于以CART为弱分类器的基于boosting的方法。这说明在多模态情感识别中, 恰当地考虑各个模态的重要性的不同是非常必要的。

## 4 结论和展望

本文提出了一种基于boosting的双模态情感识别算法, 它以CART模型为弱分类器, 能够在训练过程中自动地调整不同特征的重要性。这种算法结合了决策层融合算法和特征层融合算法的优点, 从而更好地融合双模态的信息。

在以后的工作中, 还需要考虑不同模态特征数据的可信性问题。因为在实际应用中很有可能某一模态的数据产生错误, 不应该被输入到识别算法中。如何判断获取数据的可信性是需要解决的问题之一。另外, 录制更大的数据库也是值得考虑的问题, 比如覆盖更多的情感种类或者更多的发音人等等。

## 参考文献 (References)


[1] Huang T S, Chen L, Tao H. Bimodal emotion recognition by man and machine [C]// Proc ATR Workshop on Virtual Communication Environments. Japan: 1998.

[2] Chen L S, Huang T S, Miyasato T, et al. Multimodal human emotion/expression recognition [C]// Proc the Third IEEE International Conference on Automatic Face and Gesture Recognition. Japan: IEEE Press, 1998: 366 - 371.

[3] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated prediction [J]. *Machine Learning*, 1999, **37**: 297 - 336.

[4] Cowie R, Dougas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction [J]. *IEEE Signal Processing Magazine*, 2001, **18**: 33 - 80.

[5] Fasel B, Luttin J. Automatic facial expression analysis: A survey [J]. *Pattern Recognition*, 2003, **36**(1): 259 - 275.

[6] Silva D, Miyasato T, Nakatsu R. Facial emotion recognition using multi-modal information [C]// Proc International Conference on Information and Communications Security. Singapore, 1997: 397 - 401.

[7] Chen C Y, Huang Y K, Cook P. Visual/Acoustic emotion recognition [C]// Proc International Conference on Multimedia and Expo. Amsterdam, Netherdam, 2005: 1468 - 1471.

[8] Busso C, Deng Z, Yildirim S, et al. Analysis of emotion recognition using facial Expressions, Speech and Multimodal Information [C]// Proc the 6th Intl Conf on Multimedia Interfaces. State College, PA, USA, 2004: 205 - 211.

[9] Picard R W. Affective Computing [M]. Cambridge, Massachusetts; London, England: The MIT Press, 1997.

[10] Ekman P, Huang T S, Sejnowski T J, et al. Final report to NSF of the planning workshop on facial expression understanding [R]. Human Interaction Lab, Univ California, San Francisco, 1993.

[11] Lee C M, Yildirim S, Bulut M, et al. Emotion recognition based on phoneme classes [C]// Proc International Conference on Spoken Language Processing. Jeju, Korea, 2004: 889 - 892.

[12] Schuller B, Rigoll G, Lang M. Hidden Markov model based speech emotion recognition [C]// Proc ICASSP. Washington, DC, USA: IEEE Computer Society, 2003, **2**: 401 - 404.

[13] Tian Y, Kanade T, Cohn J F. Recognizing action units for facial expression analysis [J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2001, **23**(2): 97 - 115.

[14] Tao J H, Kang Y G. Features importance analysis for emotion speech classification [C]// Proc the 1st International Conference on Affective Computing and Intelligence Interaction. Heideloerg Berlin: Springer, 2005: 449 - 457.